DECLARATION OF KONSTATINOS PSOUNIS, PHD

Redacted Version of Document Sought to be Sealed

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA OAKLAND DIVISION

CHASOM BROWN, et al., individually and on behalf of all similarly situated,

Plaintiffs,

vs.

GOOGLE LLC,

Defendant.

Case No. 4:20-cv-03664-YGR-SVK

DECLARATION OF KONSTANTINOS PSOUNIS, PHD

TABLE OF CONTENTS

Page

BACKGROUND AND QUALIFICATIONS

1. I have been retained by Google to analyze and respond to certain opinions proffered by Plaintiffs' retained experts in support of Plaintiffs' Response to Google's Submission Regarding Order to Show Cause (Dkt. 833-1) ("Plaintiffs' Response").

2. I am a Professor and Associate Chair of Electrical and Computer Engineering and Professor of Computer Science at the University of Southern California, where I teach courses on networked distributed systems, probability and information theory. I joined the University of Southern California in 2003, after completing my PhD at Stanford University as a Stanford Graduate Fellow. I have published more than 100 technical papers in the field of networked distributed systems, which have been cited tens of thousands of times. I have also been awarded numerous grants and significant funding from the government and industry leaders to advance these fields. As a result, I have been named an Institute of Electrical and Electronics Engineers (IEEE) Fellow, the highest grade of membership, and a Distinguished Member of the Association of Computing Machinery (ACM) for my contributions to the theory and practice of networked, distributed systems. Attached hereto as Exhibit A is a true and correct copy of my curriculum vitae.

3. My professional career has spanned more than 20 years. As set forth in Exhibit A, I have extensive experience in the field of networked distributed systems, including the Internet and the world wide web, content-delivery networks, data centers and cloud computing, and wireless mobile networking systems. Throughout my career, I have analyzed, designed, and developed efficient, privacy-preserving networked distributed systems for the Internet and the Web, content-delivery networks, data centers and cloud systems, and wireless mobile networking systems. As such, I have acquired expertise in the analysis and development of those systems. I have also been the faculty in charge of the entire networking curriculum at the Electrical and Computer Engineering department at USC for more than a decade and teach networking classes as well as probability theory classes which cover entropy, statistics, and other related concepts to graduate students yearly. In my analysis of networked distributed systems and my associated technical publications I regularly use probabilistic and statistical approaches including random sampling.

OPINIONS

4. I have reviewed the declaration of Mr. Chris Thompson (Dkt. 833-3) ("Thompson Decl.") and the declaration of Mr. Jay Bhatia (Dkt. 833-4) ("Bhatia Decl."), Plaintiffs' Response (Dkt. 833-1) and the briefs and declarations and underlying materials Google submitted in response to the Court's Order to Show Cause (Dkts. 797-3 to 797-22). As explained further below, I have reached the following opinions:

5. **Opinion 1:** Mr. Thompson's opinion that "Google joins authenticated and unauthenticated data" because authenticated data and unauthenticated data reside in the same log, Thompson Decl. ¶ 39 and Ex. A, is without basis.

a. Mr. Thompson appears to understand "join" and "joining" of data as "storing authenticated data and unauthenticated data in the same log." Thompson Decl. ¶ 35.

b. Mr. Thompson's use of "join" and "joining" runs counter to common usage in data analytics and computer science, and the well-established definitions of "join" and "joining" data used in more than 70 years of technical literature.

c. Mr. Thompson's unusual understanding of "joining" also conflicts with Plaintiffs' other experts' statements and his own prior testimony.

d. Mr. Thompson's assertion that my opinion that "the [redacted] [log] 'does not join authenticated data with unauthenticated data' . . . is misleading and incomplete," Thompson Decl. ¶ 34, is incorrect for the same reasons.

e. Mr. Thompson has not presented any evidence that Google joins individual unauthenticated records with individual unauthenticated records. Based on all of the evidence I have reviewed,[1] Google does not join authenticated and unauthenticated data.

f. Mr. Thompson's Exhibit 9 is misleading and does not show Google joining authenticated and unauthenticated data.

[1] *See supra* ¶ 4; Dkt. 797-21 ("November 30, 2022 Declaration") ¶ 3; Rebuttal Report Appendix H.

6. **Opinion 2:** Mr. Thompson's opinion that the existence of "a log which contains both 'unauthenticated' and 'authenticated' data . . . makes it even easier to join private browsing data with users' Google accounts" Thompson Decl. ¶ 37, is without basis.

a. Mr. Thompson's opinion that it is easier to join data stored in the same log contradicts his opinion that data is already joined when it resides in the same log.

b. As I explained in my June 7, 2022 Rebuttal Report (Dkt. 659-10) ("Rebuttal Report"), (i) Google maintains policies and technical barriers to prevent the data from being joined, and (ii) applying established principles of information theory to common networked distributed systems demonstrates that the data can not be reliably joined by IP address and User Agent.

c. The existence of a log that contains both authenticated and unauthenticated data does not resolve any of the issues I identified in my Rebuttal Report that make Plaintiffs' proposed fingerprinting methodology unreliable.

d. Mr. Thompson has not reviewed the source code that formed the basis for the opinions in my November 30, 2022 Declaration or proposed any way to validate his proposed methodology for joining authenticated and unauthenticated records (by, *e.g.*, proposing a way to eliminate false matches).

e. Mr. Thompson fails to consider the likelihood (or even the possibility) of a false match and ignores common scenarios in which it will be impossible to determine what individual person's site activity data is associated with an IP address and a User Agent.

f. Mr. Thompson's proposed use of "IP address, User agent, maybe_chrome_incognito, [redacted], [redacted], [redacted] and [redacted]" to join records within [redacted] [redacted] suffers from the same infirmities described in the rebuttal report I submitted in this litigation on June 7, 2022. *See* Dkt. 659-10 ("Rebuttal Report") at Opinions 1, 2, 5, 6, 7, 8, 9, 10, 11, 12, and 13.

7. **Opinion 3:** Mr. Bhatia's opinion that I did not have a "reasonable basis to say, 'the coding of the [[redacted]] log prohibits it from ever joining authenticated with unauthenticated data,'" Bhatia Decl. ¶ 26, is incorrect:

a. Mr. Bhatia's understanding of the term "joining" also runs counter to common usage in data analysis and computer science, along with more than 70 years of technical research.

b. The Source Code that I reviewed provides a sufficient basis to conclude that the coding of the [redacted] Log prohibits it from ever joining authenticated data with unauthenticated data.

c. Mr. Bhatia's opinion that records in [redacted] [redacted] are joined or merged sequentially by time is incorrect.

8. **Opinion 4:** Plaintiffs' assertion that a one percent random sample is invalid, Dkt. 833-1 at 11-12, is without basis.

a. A sample is not invalid merely because it is based on 1% of random sampling.

b. In a large population of, for example, billions of log records, a 1% random sample allows for analysis of tens of millions of records.

Opinion 1: Mr. Thompson's opinion that "Google joins authenticated and unauthenticated data" because authenticated and unauthenticated data are stored in the same log, Thompson Decl. ¶ 39 and Ex. 9, is contrary to the industry and academic definition of "joining."

9. Although he does not expressly define the term "joining," Mr. Thompson appears to understand "joining" as storing different records from two separate logs in a single log without linking any records together. That notion of "joining" is contrary to how "joining" is commonly defined and used in data analysis and computer science. Mr. Thompson's use of "joining" does not align with over 70 years of technical literature addressing this concept, commonly used "join" functions in the Python and SQL programming languages, and my experience in the field.

A. Mr. Thompson's Understanding of "Joining" Runs Counter to Common Usage in Data Analytics and Computer Science, and More Than 70 Years of Technical Research.

10. Mr. Thompson is correct that I "define[] 'joining' to mean that 'a shared key (or any common data point) was used to associate or combine unauthenticated private browsing data at issue with an individual's Google account.'" Thompson Decl. ¶ 35. I also explained that "[f]or the logs in question, authenticated and unauthenticated data would be considered 'joined' if a log shows that a shared key (or any common data point) was used to associate or combine unauthenticated private browsing data at issue with an individual's Google account." Dkt. 797-21 ("November 30, 2022 Declaration").

11. The definition of "joining" in my November 30, 2022 Declaration conforms to well-established definitions from a long line of technical literature defining "joining" or "linking" of datasets as merging records from multiple datasets into combined records (*i.e.*, records that contain information from more than one "input" dataset) via the use of a common join key, and it aligns with the definition employed by researchers in this field for decades.[2] Common database programming languages like Python and SQL define "join" and "joining" in the same way.

[2] *See, e.g.*, Rob Hall and Stephen E. Fienberg, "Privacy-Preserving Record Linkage," 2010 Int'l Conf. on Priv. in Stat. Databases, https://www.cs.cmu.edu/~rjhall/linkage_survey_final.pdf at 1 (Sept. 2010) ("Record linkage is an historically important statistical problem arising when data about some population of individuals, is spread over several files. Most of the literature focuses on the two file setting. *The record linkage goal is to determine whether a record from one file corresponds to a record of a second file, in the sense that the records describe the same individual*." (emphasis added)); Ahmed K. Elmagarmid, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf, 19:1 at 1 (Jan. 2007) ("[T]he construction of a comprehensive view of . . . data [from multiple data sets] consists of linking—in relational terms, joining—two or more tables *on their key fields*." (emphasis added)); Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," Journal of Amer. Statistical Assoc. 64:328, https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf at 51 (Dec. 1969) ("The necessity for comparing the records contained in a file L_A with those in a file L_B in an effort to determine which pairs of records relate to the same population unit is one which arises in many contexts, most of which can be categorized as either (a) the construction or maintenance of a master file for a population, or (b) *merging two files in order to extend the amount of information available for population units represented in both files*." (emphasis added)); Howard B. Newcombe and James M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," Comm. of the Assoc. for Computing Machinery 130:3381, https://dl.acm.org/doi/pdf/10.1145/368996.369026 at 563 (Oct. 16, 1959) ("Linkage of a *pair of records* relating to a particular individual or family involves two steps: first, a searching operation in which potentially linkable records are brought together for scrutiny, followed by a detailed comparison to decide whether the person or persons referred to on each are in fact the same." (emphasis added)); Halbert L. Dunn, "Record Linkage," Am. Journal of Public Health, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1624512/pdf/amjphnation00640-0051.pdf, at

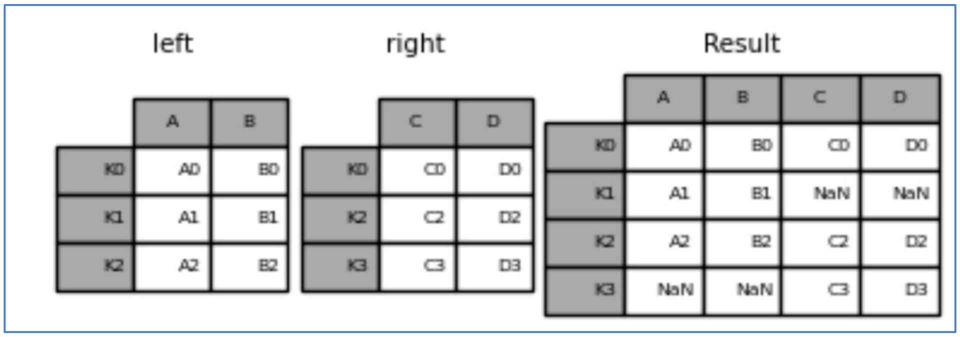
12. Python is one of the most widely used programming languages for data analysis today. In Pandas, which is a popular open source Python library for data analysis initially released in 2008,[3] the "join" function is used to combine records from two or more database-style "DataFrames" based on a common key between them, whereas the default "concat" function is used to concatenate records from two or more database-style "DataFrames" by merely listing them one after the other without considering the presence of any common key between them.[4] The following figures illustrate the distinction between these three functions in practice:

1414 (Dec. 1946) (describing "record linkage" as linking certain statistics with "other facts about the same individuals").

[3] *See* https://pandas.pydata.org/about/.

[4] The .join() function is using the more general merge() function. The concat() function concatenates along the rows axis by default. For more details about these functions see "Merge, join, concatenate and compare," https://pandas.pydata.org/docs/user_guide/merging.html.

Figure 1 – Python .join() Function[5]



left

	A	B
K0	A0	B0
K1	A1	B1
K2	A2	B2

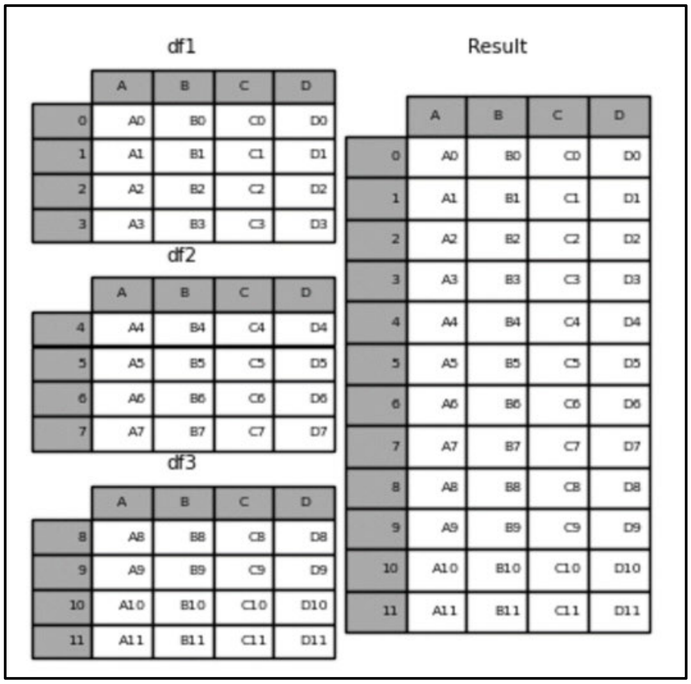
right

	C	D
K0	C0	D0
K2	C2	D2
K3	C3	D3

Result

	A	B	C	D
K0	A0	B0	C0	D0
K1	A1	B1	NaN	NaN
K2	A2	B2	C2	D2
K3	NaN	NaN	C3	D3

Figure 2 - Python Default concat() Function



df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

Result

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

13. As the figures above illustrate, "joining" records in Python causes input records that share a common key to be combined into a single record in the output table, and thus any "joined" output records include data from both input records. By contrast, Python's default "concat" command merely adds records to the output table without generating any "joined" records that contain information from more than one input.

[5] The rows of the "left" and "right" input DataFrames with common keys K0 and K2 are joined. (Shown join result with parameter how="outer".).

14. Similarly, in Structured Query Language ("SQL")—a programming language commonly employed for database management and analysis—the JOIN statement (also referred to as INNER JOIN) is used to combine data or rows from two or more tables based on a common field between them.[6]

15. By contrast, according to Mr. Thompson's use of "joining", two datasets would be considered "joined" merely if they were copied and pasted into the same table. "Joining" or "linking" records as it is defined in technical literature (and used in common programming languages) has been a fruitful subject for extensive inquiry because the use of common keys to "join" datasets is a complex subject with broad applications in data analysis.[7] However, copying and pasting two tables into another table is trivial and does not generate significant research questions. Mr. Thompson's use of the term "joining" does not align with over 70 years of technical literature addressing this concept, commonly used "join" functions in the Python and SQL programming languages, and my experience in the field.

B. Mr. Thompson's Understanding of the Term "Joining" Conflicts With Plaintiffs' Other Experts' Statements and His Own Prior Testimony.

16. Mr. Thompson's use of "join" and "joining" also conflicts with Plaintiffs' Expert Mr. Hochman's testimony. *See* Hochman Vol. I Dep. Tr. 95:10-14 ("I think in the report, I've demonstrated how the fingerprinting information can be used to *join records from different logs* and to reidentify private browsing activity." (emphasis added)); *id.* at 101:14-102:1 (describing joining as "match[ing] up" records, where you can "look at the fingerprinting information and see if you've got a correct join or maybe a *false match*") (emphasis added); Hochman Opening Report (Dkt. 608-12) ¶ 237 (discussing the purported feasibility of "join[ing] a user's *private browsing activities* on non-Google websites *with the user's Google account identity*[.]" (emphasis added)).

17. Mr. Thompson's conception of "joining" also conflicts with Plaintiffs' Expert Mr. Schneier's definition of "joinability":

[6] *See generally* Mark Reed, *SQL: 3 books in 1 - The Ultimate Beginner, Intermediate & Expert Guides To Master SQL Programming Quickly with Practical Exercises*, Ch. 4 (2022).

[7] Mr. Thompson's proposed definition also leads to an absurd result here because (i) if any records that are stored in the same log file are "joined," then (ii) it follows that all of the records stored in any one of Google's logs would be considered to be "joined" with all other records in the same log.

Joinability is the process of linking two data sets. One definition:

> *Joinability measures whether data sets are linkable by unexpected join keys*. Sometimes it is necessary to retain multiple data sets with different ID spaces. In those cases data custodians should avoid linking the two data sets to respect the choices of users who maintain separate identities. As an example, consider a website that can be used either signed-in or signed-out. A user may choose to use the website signed-out to separate activities from their signed-in identity. If the website operator maintains data sets about activities of both signed-in and signed-out users, it might accidentally include granular information (e.g. web browser user agent) in both data sets that *could allow the signed-in and signed-out identities to be linked*. In that case, we would say that the identities in the two data sets are joinable.

Schneier Report (Dkt. 608-7) ¶ 153 (quoting Pern Hui Chia, et al., "KHyperLogLog: Estimating Reidentifiability and Joinability of Large Data at Scale," Proceedings of the IEEE Symposium on Security and Privacy, https://milinda-perera.com/pdf/CDPSLDWG19.pdf (2019) (emphasis added); *see also id.* at ¶ 205 ("[E]ven if Google is not *building user profiles across signed-in and signed-out data*, Google's decision to collect and log this data creates the potential for data to be *joined in this way*." (emphasis added)); Schneier July 18, 2022 Dep. Tr. 199:18-20 ("In general, the methods of joining involve taking the two data sets, and *through a variety of different correlation techniques, matching them*." (emphasis added)). By contrast, Mr. Thompson opines that data records can be described as "joined" without "matching" any records. Mr. Thompson's opinion conflicts with Mssrs. Hochman and Schneier, and it runs counter to common usage in the field.

18. Mr. Thompson's understanding of "join" and "joining" also conflicts with his prior testimony, in which he described "joining" of data as "how do we connect and how do we look at those logs and analyze those logs to potentially *connect the data within them*," Apr. 21, 2022 Hrg. Tr. 41:16-23 (emphasis added), and he proposed that "there are . . . ways to link [these] columns [showing authenticated, pseudonymous, and Google analytics-keyed data]," *id.* at 78:20. As to the "ways to link" authenticated and unauthenticated data, Mr. Thompson testified that authenticated and unauthenticated data could be joined by (i) using encrypted pseudonymous identifiers from a user who signs into Google while in Incognito mode to link the user's signed-in activity with signed-out activity, *id.* at 78:1-11; and (ii) using third-party identifiers to join profiles across different sessions, *id.* at 87:19-23.

19. Although Mr. Thompson's assertions regarding the reliability of Plaintiffs' proposed fingerprinting methodology are incorrect for the reasons discussed in my second opinion and in my Rebuttal Report (Opinions 1, 3, 5-7, 9-12, 13; Appendices E and F), his prior testimony aligns with the well-established technical definitions discussed above (*i.e.*, "joining" data requires matching individual records from two datasets and combining them into a single record via the use of a common key). Mr. Thompson's new understanding—where data is purportedly "joined" merely because it is stored in the same log—does not.[8]

20. This difference is not just semantic. Merely adding records to the same log does not join the data in those records because it does not create any linkages between individual records. In other words, even if authenticated and unauthenticated records resulting from the same user's browsing activity are contained in the same log, none of the unauthenticated records are associated with the user's identity because these individual records are not "joined" or "linked" to the authenticated records. In practice, this means that querying [redacted] for an authenticated GAIA identifier will not return results for any unauthenticated data.

C. Mr. Thompson Has Not Reviewed the Source Code That Formed the Basis for My November 30, 2022 Declaration or Proposed Any Way to Validate His Proposed Methodology or Eliminate False Matches.

21. In my November 30, 2022 Declaration, I explained the basis for my conclusion that [redacted] does not join authenticated data with unauthenticated data:

> In reaching these opinions, I reviewed the following information: Google employee declarations and exhibits thereto that Google is submitting concurrently with this declaration, source code related to the [redacted] log, Plaintiffs' August 4, 2022 (Dkt. 655-1) and August 25, 2022 (Dkt. 707-1) briefs

[8] I understand that Plaintiffs' counsel has also applied a definition of "join" or "link" that conflicts with Mr. Thompson's new conception in statements to the Court. *See* June 2, 2021 Hrg. Tr. 13:24-14:3 (Mr. Mao: "From documents which we have seen in *Brown*, our understanding is that the GAIA I.D.'s, or authenticated logged in users, are often *paired or linked, or at least in our definition merged, but at least correlated and confirmed against the unauthenticated data*." (emphasis added)); Nov. 4, 2021 Hrg. Tr. 15:21-25 (Ms. Bonn: "Google's own documents show that, and I'm showing one example, *IP addresses and locations are passed into search logs and are semi-unique fingerprints which can join keys for linking Incognito Zwieback activity to GAIA, users' accounts*." (emphasis added)); *id.* at 47:10-14 (Ms. Bonn: "So using things like their IP address, their user agent string or other identifiers. That is going to give us so much more insight into how these systems work at Google, how *these fields can be joined*, maybe how they can't, so that key can then formulate an intelligent query." (emphasis added)).

> related to their request for supplemental sanctions, Exhibit B attached to this declaration, and publicly-available technical literature regarding certain code functions. Google provided me with all information I asked for to enable me to render the opinions in this declaration.

November 30, 2022 Decl. ¶ 3.

22. To analyze my opinion on this log, I would expect Mr. Thompson to review the same source code to determine whether records within [redacted] are joined.

23. To support his opinion that data is joined, I would also expect Mr. Thompson to explain how the purported joins in this log could be verified to eliminate any false matches. Technical literature explains that "[l]inkage of a pair of records relating to a particular individual or family involves two steps: first, a searching operation in which potentially linkable records are brought together for scrutiny, *followed by a detailed comparison to decide whether the person or persons referred to on each are in fact the same*." Howard B. Newcombe and James M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," Comm. of the Assoc. for Computing Machinery 130:3381, https://dl.acm.org/doi/pdf/10.1145/368996.369026 at 563 (Oct. 16, 1959) (emphasis added). Plaintiffs' expert Mr. Hochman also noted that this second verification step is essential to joining records. Hochman Vol. I Dep. Tr. 101:14-102:1 (describing joining as "match[ing] up" records, where you can "look at the fingerprinting information and see if you've got a correct join or maybe a false match"). Mr. Thompson does not identify any process for this verification.

D. Mr. Thompson's Exhibit 9 Does Not Show Google Joining Signed-Out Private Browsing Data With Users' Google Accounts.

24. Exhibit 9 to Mr. Thompson's declaration does not show any joining or linking of records as those terms are used in technical literature and in the field. Exhibit 9 merely shows multiple unique records added to the same log—not "joining" those separate records via the use of a shared join key. For Exhibit 9 to illustrate "joining" of records, it would need to show the "match.com" and "eharmony.com" private browsing records being combined with authenticated records into a single record via a shared key. In my opinion, storing separate records in the same log

without matching or correlating any of those individual records does not constitute "joining" as that term is used in technical literature and in the field.

25. Moreover, Mr. Thompson's Exhibit 9 appears to depict all of the authenticated and unauthenticated records in a log being generated from the same individual's website visits with the same IP address and User Agent. In my opinion, this is highly misleading because Google's logs will obviously contain many more records (by orders of magnitude) and as discussed *infra* in Opinion 2, and in my Rebuttal Report (Opinions 1, 2, 5, 6, 7, 8, 9, 10, 11, 12, and 13; Appendices E and F), there are many common instances where an IP address and User-Agent will not be unique to an individual (*e.g.*, when multiple individuals in the same household share a desktop computer, or where employees access the internet via a shared company-supplied VPN and use a company-managed browser, or an Internet Service Provider (ISP), or a company or an organization that dynamically allocates IP addresses using DHCP). In my opinion, Mr. Thompson's depiction of one individual in Exhibit 9 is based on a faulty assumption that the same person used the laptop depicted in Mr. Thompson's diagram in signed-out private browsing mode at 8:34 and in signed-in regular mode at 8:40. As Mr. Hochman testified, to actually verify whether unauthenticated data was generated by a specific person would require individual interviews and investigation.[9]

26. For the reasons discussed below in Opinion 2 and in my Rebuttal Report (Opinions 1, 2, 5, 6, 7, 8, 9, 10, 11, 12, and 13; Appendices E and F), the data illustrated in Exhibit 9 does not provide sufficient information to reliably conclude that signed-out private browsing mode records stem from the same user that generated the signed-in browsing records.

[9] *See* Hochman Dep. Tr. Vol. I 95:20-96:14 ("Q. And your opinion is that the private browsing information at issue in this case could be reidentified through fingerprinting techniques, correct? A. Well, in fact, I think I know that it has been -- private browsing information has been reidentified. There are examples of it happening. I think -- I spoke with Dave Nelson and I think – and I read his deposition. And I think you've heard him say that the FBI has gone and arrested people who said, How did you find me? I was in incognito mode, or I was in private mode. And so we know that that's happened. There are instances where -- where it actually has happened."); *id.* at 141:21-142:5 ("Q. How would you establish whether an IP address has successfully uniquely identified a user? A. I mean, one thing I would point you to is -- is how often law enforcement goes and seeks this information and what their success rate is in identifying people, okay? That's been written up."); *id.* Vol. II 475:8-24 ("Once you narrow things down that much [to 1000 possible matches], in any sort of case where somebody is trying to identify a user, and I've said this before in prior answers, there's always some additional information available. There are always some circumstances around that, some other things that are known, and those can be used in a process of elimination to whittle down any small group down to identify the individual. And I know that Dave Nelson will come forward and offer to testify that the FBI does that all the time, that they're not hindered by this scenario of multiple people in a household.").

Opinion 2: Mr. Thompson's opinion that the existence of "a log which contains both 'unauthenticated' and 'authenticated' data . . . makes it even easier to join private browsing data with users' Google accounts" Thompson Decl. ¶ 37 is without basis.

27. Mr. Thompson opines that the existence of "a log which contains both 'unauthenticated' and 'authenticated' data . . . makes it even easier to join private browsing data with users' Google accounts particularly in light of the fact it is now clear that there is at least one Google log that contains the following fields for both unauthenticated and authenticated (i.e., GAIA data): IP address, User agent, maybe_chrome_incognito, [redacted], [redacted], [redacted] and [redacted]." Thompson Decl. ¶ 37. Mr. Thompson further asserts that "[a]nyone seeking to join the data could easily do so by way of matching data all within the same log." *Id.*

28. Mr. Thompson's opinions are internally inconsistent because he simultaneously claims (incorrectly) that (i) "Google joins authenticated and unauthenticated data" by "storing authenticated and unauthenticated records in the same log," Thompson Decl. ¶¶ 35, 39; and (ii) "a log which contains both 'authenticated' and 'unauthenticated' data . . . makes it even *easier* to join private browsing data with users' Google accounts," *id.* ¶ 37 (emphasis added). If the former is correct (it is not), then there is no need to make joining "easier" because the records would already be joined (they are not).

29. Mr. Thompson's opinion appears to be based on the premise that it is purportedly "eas[y]" to "join private browsing data with users' Google accounts." Thompson Decl. ¶ 37. As I explained in my Rebuttal Report and below, that is a false premise because (i) the fields that Mr. Thompson identifies are not sufficiently stable and unique to reliably identify the specific device from which data was received or the individual that was browsing (*see infra* Sections A-D); and (ii) Google employs myriad technical and policy restrictions to prevent joining of signed-out private browsing data with users' Google accounts (*see infra* Section E). For a more detailed description, *see generally* Rebuttal Report Opinions 1, 3, 5, 6, 7, 9, 10, 11, 12, 13; Appendices E and F.

A. IP Addresses are Neither Unique nor Static

30. As to IP addresses, neither IPv4 nor IPv6 addresses are unique or static, and there is no one-to-one mapping between either type of IP address and individual users. An IPv4 address is

a set of four numbers, each ranging from 0 to 255, assigned to an internet-connected device or devices. Each address consists of 32 bits, yielding a total of 2^32 possible addresses. With the increase of the number of devices per person and the people using the Internet, these IP addresses were not enough and a newer version of the IP protocol, IPv6, defined new IP addresses each consisting of 128 bits. Adoption of IPv6 worldwide is still less than 40 percent and currently at 46.74 percent in US,[10] and thus at least 50 percent of the IP addresses in the US are IPv4 addresses. Unless otherwise stated, I use the term IP address to refer to both IPv4 and 1Pv6 addresses.

31. It is rarely the case that a device has a unique, static IP address. For example, devices inside the home of a family of four who own a total of eight devices combined will all share the same external IP address, which is then internally translated to eight distinct local IP addresses using the so-called NAT (Network Address Translation) protocol.[11] There is no way to tell which of the eight devices is the one that accesses a website, if all one has as an identifier is the external IP address of the device.

32. As another example, employees of a large company who are working from home are connected to their company via what is called a VPN (Virtual Private Network). For the vast majority of VPN services, if any of these employees access a website, the external IP address will be the same for each employee and will equal the IP address of the VPN server. Again, there is no way to know which of the hundreds of potential devices connected to the VPN server is the one that accessed the website, if the only identifier is the external IP address.

33. As one more example, consider a company that assigns dynamic rather than static IP addresses to devices inside the company's network, using the so-called DHCP (Dynamic Host Configuration) Protocol.[12] Note that due to the flexibility of DHCP over static IP addresses, it is the preferred method to assign IP addresses to end-user devices across large organizations.[13] Identifying a device by its IP address when DHCP is used is very unreliable because one cannot know for sure which device has a specific dynamic IP among the possible set of IP addresses at any given time,

[10] GoogleIPv6, https://www.google.com/intl/en/ipv6/statistics html (last visited January 30, 2023).
[11] J.F. Kurose & K.W. Ross, *Computer Networking: A Top-Down Approach*, Ch. 4 (8th ed. 2020).
[12] *Id.*
[13] Microsoft, "Dynamic Host Configuration Protocol (DHCP)," https://perma.cc/7N5L-QKCQ (July 29, 2021).

since the assignment of IP addresses to devices changes dynamically. For example, in a company with 200 devices and 256 IP addresses, any device may have any of the 256 IP addresses at any given time. There is no way to know for sure which of the 200 devices is the one that accesses a website, if the only identifier is the external IP address of the device that accesses the website.

34. One more reason why a device may not have a unique, static IP address is the use of Onion Routing, known to most as Tor.[14] Tor directs internet traffic through an overlay network consisting of thousands of relays and uses a series of layered nodes to hide the IP address of a device. In practice, a device that uses Tor to access a website cannot be identified by the observed IP address, which is the IP address of the so-called exit node.

35. Because the number of possible IPv6 addresses is so vast, there is a misconception that an IPv6 address always identifies a single device. This is inaccurate. First, VPN services that support IPv6 hide both IPv4 and IPv6 device addresses in a seamless fashion.[15] Second, IPv6 NAT products (see, for example, IPv6 NAT functionality within Junos OS of Juniper Networks) not only support address translation between IPv4 and IPv6 addresses, but also between IPv6 hosts. In particular, NAT between IPv6 hosts is done in a similar manner as IPv4 NAT, and, in this case, a single, public IPv6 address may correspond to a large number of private IPv6 addresses. In addition, to address privacy concerns with dynamic IPv6 addresses assigned by DHCPv6, privacy extensions allow clients to use random lower 64bit IIDs.[16] For the upper 64bit IID, providers employ prefix rotation via what is known as temporary mode DHCPv6,[17] made possible thanks to the sheer number of IPv6 addresses which has resulted in single residential customers often having more IPv6 addresses assigned than the entire IPv4 address space.[18] The combined result of the above random selection of IPv6 address bits, together with the large IPv6 address space, has resulted in many IPv6

[14] *See* Tor Project, https://www.torproject.org/about/history/ (last visited January 30, 2022).

[15] *See* NordVPN, https://nordvpn.com/features/hide-ip/ (last visited January 30, 2022); Surfshark, https://surfshark.com/use-cases (last visited January 30, 2022).

[16] *See* T. Narten, R. Draves, and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6," RFC 4941 (Draft Standard), https://www.rfc-editor.org/rfc/pdfrfc/rfc4941.txt.pdf (Sept. 2007).

[17] *See* Tomek Mrugalski, et al., "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)," RFC 8415 (Proposed Standard), https://www.rfc-editor.org/rfc/pdfrfc/rfc8415.txt.pdf (Nov. 2018).

[18] E. Rye, R. Beverly, and K. Claffy, "Follow the Scent: Defeating IPv6 Prefix Rotation Privacy," Proceedings of ACM Internet Measurement Conference (IMC), https://arxiv.org/pdf/2102.00542.pdf (Nov. 2-4, 2021).

addresses being used only once (*see*, for example, data from a large CDN[19] which found that more than 90 percent of IPv6 addresses appear only once in a long-running data collection campaign).[20] Additionally, Tor has recently added support for IPv6 addresses.[21]

36. Additionally, Google applies "IP address redaction" to "[a]ll logs (including [redacted] [redacted]), except for logs with a documented [redacted] exception to keep PII longer than [redacted]."[22]

37. As explained in Appendix F to my Rebuttal Report, analysis of data Google produced pursuant to the Special Master process supports my conclusion that an IP address cannot reliably identify a device. This analysis showed that out of the 4,945 distinct IP addresses I evaluated, there are 159 that have multiple GAIAs associated with them, and three of these IP addresses, specifically [redacted], [redacted], and [redacted], have multiple GAIAs that correspond to more than one plaintiff.[23]

B. User Agent, [redacted]*,* [redacted]*,* [redacted]*, and* [redacted] *Are Neither Unique Nor Static.*

38. A User Agent string (UA) contains information about the type of the browser (e.g. Chrome, Edge, Mozilla, Safari), the version of the browser, and the operating system over which the browser is running (e.g. Windows, macOS, iOS, Linux).[24] It is used to identify the type and version of the browser and the operating system such that the behavior and content of web browsing can be customized accordingly. It should be evident that millions of users share the same UA. It should be equally evident that some UAs are more common than others, for example, recent versions of popular web browsers running on top of popular operating systems are very common across devices. In fact, a recent study, which collected UAs from an Internet measurement company over

[19] D. Plonka and A. Berger, "Temporal and Spatial Classification of Active IPv6 Addresses," Proceedings of ACM Internet Measurement Conference (IMC), https://conferences2.sigcomm.org/imc/2015/papers/p509.pdf (Oct. 28-30, 2015).

[20] While the use of one-time IPv6 addresses is prevalent among non-mobile devices, cellular providers may assign static IPv6 addresses to their clients.

[21] The State of IPv6 support on the Tor Network, https://blog.torproject.org/state-of-ipv6-support-tor-network/ (last visited January 30, 2023).

[22] GOOG-BRWN-00029002, at -002 ("Google partially redacts IP addresses within [redacted] of collection . . . The current implementation is to remove the lower 8 bits (keeping 24) from IPv4 addresses, and remove the lower 80 bits (keeping 48) from IPv6 addresses.").

[23] *See* Dkt. 659-10 Appendix F: IP Address + User Agent Data Analysis.

[24] MDN Plus, "User-Agent," https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/User-Agent (last visited January 30, 2022); Chrome Developers, "User-Agent Strings," https://developer.chrome.com/docs/multidevice/user-agent/ (updated Nov. 9, 2021).

the course of two years, found that the top 10 most popular UAs correspond to 26 percent of daily traffic.[25] With about 50 billion total HTTP requests per day in the collected data, this implies that more than one billion daily HTTP requests correspond to the same UA.[26] If a device makes on average 10–100 HTTP requests during a day, this would imply that tens of millions of devices from which data has been collected share the same UA.

39. Mr. Thompson also fails to account for the possibility that a User Agent value can be changed by the user via the use of, for example, a "user agent switcher plugin."[27] Mr. Thompson does not account for the fact that "UA strings are notoriously misleading,"[28] nor does he explain whether the records described in Exhibit 9 are associated with Chrome, Safari, or Mozilla.

C. The Existence of Shared Devices Make Plaintiffs' Proposed IP + UA Fingerprinting Method Unreliable.

40. In my opinion, the reality of device sharing makes Mr. Thompson's proposed use of a combination of IP address and user agent to tie signed-out private browsing mode data to users' account unreliable. There is a significant body of research—including research published by Plaintiffs' expert Mr. Schneier—demonstrating that sharing of devices by more than one user is commonplace.[29] Mr. Schneier's research in this area notes that "[p]eople living in the same

[25] J. Kline, P. Barford, A. Cahn, and G. Sommers, "On the structure and characteristics of user agent strings," ACM Sigcomm, https://conferences.sigcomm.org/imc/2017/papers/imc17-final253.pdf (Nov. 1-3, 2017).

[26] 26 percent of the 50B HTTP requests is more than 12.5B HTTP requests. Hence, on average, each of the 10 most popular UAs corresponds to more than 1.25B HTTP requests.

[27] *See, e.g.*, Rebuttal Report ¶ 123 (discussing user agent switcher plugin) (citing Jonathan Hochman - Twitter, https://twitter.com/Jehochman/status/1153277584542711808 (last visited Feb. 8, 2023) (recommending use of user agent switcher plugin to access old twitter interface: "Set it to IE11").

[28] https://developer.mozilla.org/en-US/docs/Web/HTTP/Browser_detection_using_the_user_agent (last visited February 8, 2023); *see also id.* ("Using the user agent to detect the browser looks simple, but doing it well is, in fact, a very hard problem.").

[29] *See, e.g.*, Tara Matthews, et. al., "'She'll just grab any device that's closer': A Study of Everyday Device and Account Sharing in Households," Proceedings of the ACM Conference on Human Factors in Computing Systems, ACM, https://dl.acm.org/doi/pdf/10.1145/2858036.2858051 (2016), at 2 ("Among our key findings are that device and account sharing is common, and that mobile phones were shared as much as computers and more often than tablets."); K. Levy and B. Schneier, "Privacy threats in intimate relationships," Journal of Cybersecurity, https://academic.oup.com/cybersecurity/article/6/1/tyaa006/5849222 (2020) ("Schneier Privacy Threats"), at 10 ("[H]ouseholds are not units; devices are not personal; the purchaser of a product is not its only user."); *id.* (criticizing the "assumption . . . that devices considered 'personal' are used by only one person" because "abundant research demonstrates that this is often not the case"); A. Brush and K. Inkpen, "Yours, Mine and Ours? Sharing and Use of Technology in Domestic Environments," Proceedings of the 9th International Conference on Ubiquitous Computing, https://www.microsoft.com/en-us/research/wp-content/uploads/2007/09/brushinkpenyoursmineours.pdf (2007); B. Busse and M. Fuchs, "Prevalence of Cell Phone Sharing," Survey Methods: Insights from the Field, https://surveyinsights.org/?p=1019 (2013); H. Muller, J. Gove, and J. Webb, "Understanding Tablet Use: A Multi-Method Exploration," Proceedings of the 14th international conference on Human-computer interaction with mobile

household may share computers, phones, and other connected devices."[30] He further explains that the frequency with which devices are shared among users undercuts "[s]ystem designers['] buil[t] in assumptions about intrafamilial privacy expectations," which leads to incorrectly "treat[ing] a household as a 'unit' for purposes of information sharing."[31] Mr. Schneier urges that we should "realize that households are not units[,] devices are not personal[,] [and] the purchaser of a product is not its only user."[32]

41. In my opinion, the existence of shared devices make Mr. Thompson's proposed fingerprinting methodology unreliable because there is no deterministic method to establish whether data came from a shared device and, for shared devices, there is no reliable one-to-one mapping between (i) the combination of IP address and user agent; and (ii) individual users.

42. If one were to attempt to employ the joining methodology that Mr. Thompson proposes, it would lead to incorrect joins (*i.e.*, joining the private browsing data of User A with information keyed to User B's Google account) because there are many instances where User A and User B's browsing activity will share an identical IP address and user agent. Mr. Thompson also fails to describe how his method could account for users sharing devices and users using multiple devices at the same time.

43. Consider, for example, a desktop computer shared by a family of four. If (i) only one family member signs into a Google account on this shared device (the "Lone Signed In User"); and (ii) the IP address and user agent of the shared device is used to join records of signed-out private browsing information with records keyed to the Lone Signed In User's Google account; then (iii) records containing information about signed-out private browsing conducted by the other three family members would be joined with the Lone Signed-In User (even if he or she never personally used Incognito mode on the shared device).

44. As another example, consider three persons: John, Mary, and Peter. John and Mary live in the same house. John owns a phone and shares a home laptop with Mary. Mary owns a

devices and services, https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/38135.pdf (2012).

[30] Schneier Privacy Threats at 2.

[31] *Id.* at 10.

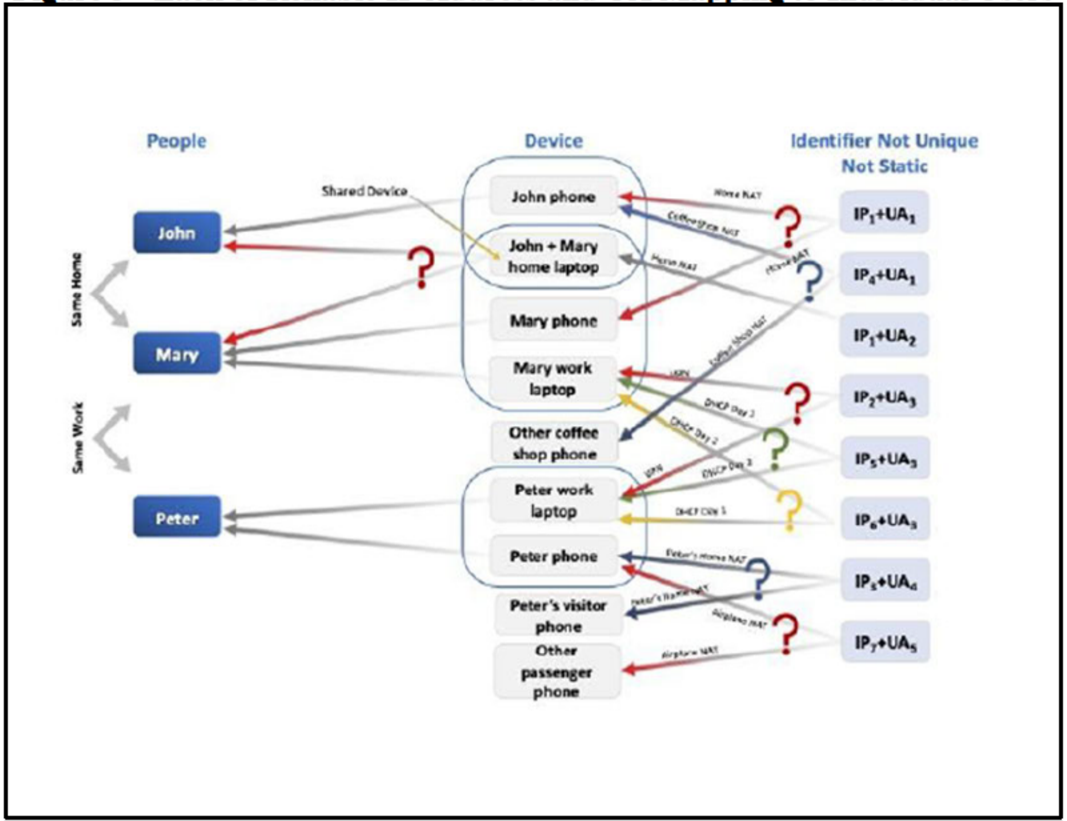
[32] *Id.*

phone, the same model as John's, and a work laptop. When at home, John's and Mary's devices connect to the home WiFi and sit behind a NAT. Mary and Peter are co-workers. Peter owns a phone and a work laptop. Sometimes they work from home and connect to their employer's network via the same VPN server. When they work in the office, their employer uses DHCP to assign IP addresses to their work laptops, and the IT department updates the software on the company's devices.

45. The following diagram shows the complexity (and impossibility) of mapping between IP+UA values, devices, and persons in this scenario. John's and Mary's phones, when at home, share the same IP+UA. John's and Mary's browsing sessions via their shared home laptop share the same IP+UA. John's and another user's phone may share the same IP+UA if they both connect to public wifi offered by a coffee shop and the other user uses the same phone model and browser as John. Mary's and Peter's work laptops, when working from home, share the same IP+UA. Mary's and Peter's work laptops, when in the office, may have the same IP+UA on different days. Peter's phone may share the same IP+UA with the phone of a visitor in his home or a passenger on the same plane, if the other person uses the same phone model and browser as Peter. In all of these cases, merely involving less than a handful of people, it is impossible to uniquely identify an individual through an IP+UA combination.[33]

[33] This example, and the lack of a reliable one-to-one mapping between IP addresses, user agents, and users are discussed in further detail in Appendix E of my Rebuttal Report.

Figure 3 - Lack of Reliable IP Address and UA Mapping to Individual Users



D. The Existence of a Log That Contains Both Authenticated and Unauthenticated Data Does Not Resolve Any of These Issues or *__Make It "Easier" to Join Signed-out Private Browsing Mode Data With a User's Google Account.__*

46. Mr. Thompson's opinion that "a user browsing in regular mode who then opened a private browsing window would have data from both contemporaneous sessions in such a way that it would make identifying their traffic through combinations of user agent and IP address trivial," Thompson Decl. ¶ 38, is also incorrect because the use of a user agent value and IP address to join data via Plaintiffs' proposed fingerprinting methodology will not reliably match data tied to a user's Google account with the same user's signed-out private browsing data for the reasons discussed above and in my Rebuttal Report (Opinions 1, 3, 5, 6, 7, 9, 10, 11, 12, 13; Appendices E and F).

47. Mr. Thompson opines that the existence of "a log which contains both 'unauthenticated' and 'authenticated' data . . . makes it even easier to join private browsing data with users' Google accounts." Thompson Decl. ¶ 37. This assertion is incorrect because merely adding records from two datasets to the same dataset does not resolve any of the issues I have described above and in my Rebuttal Report (Opinions 1, 3, 5, 6, 7, 9, 10, 11, 12, 13; Appendices E

and F). In my opinion, adding these records to the same log does not make it "easier to join private browsing data with users' Google accounts" because there is still no reliable one-to-one mapping between these identifiers and individual users—whether the records are contained in the same log or multiple different logs. In my opinion, merely adding unauthenticated records and authenticated records to the same log does not resolve any of these problems (or make them any easier to solve).

E. Mr. Thompson Fails to Account For Google's Policy and Technical Restrictions Designed to Prevent the Type of Fingerprinting That He Describes.

48. Mr. Thompson also does not account for the many policy and technical restrictions discussed in Opinions 1, 3, 4, 6, 7, and 13 of my Rebuttal Report.

49. Google has a number of policies and guidelines concerning the collection, storage, usage and deletion of data. These policies and guidelines include, for example:

- the Device/App/Browser Fingerprinting and Immutable Identifiers Policy;[34]
- the User Data Access Policy;[35] and
- the User Data Retention and Deletion Policy.[36]

50. Mr. Thompson's claim that data can be joined using IP addresses and user agent values amounts to an assertion that Google could engage in "fingerprinting" to join signed-out private browsing data with users' Google accounts. In the context of browser communications, fingerprinting refers to the use of a combination of various bits of information to probabilistically identify a browser. Probabilistic identification of a browser through fingerprinting is different from the deterministic identification of an individual through an account log-in. Actors attempting fingerprinting most commonly use network-related information such as an IP address and web browser-related information such as the type and version of the browser (i.e. User Agent).

51. Google's internal policies expressly prohibit fingerprinting unless it is to prevent spam, abuse, fraud, or other such user-beneficial usages completely unrelated to Plaintiffs' allegations.[37] Google's strict enforcement of this policy was confirmed by every Google employee

[34] GOOG-BRWN-00029326.

[35] GOOG-CABR-05455683.

[36] GOOG-CABR-00073922.

[37] *See* GOOG-CABR-04720562, at -563; GOOG-CABR-00073873, at -873 ("Device/App/Browser Fingerprinting or Immutable Identifiers must not be used by any Google products or services for the purposes of: Tracking user behavior, including: Ad measurement and prediction[,] Ad targeting[,] [and] Recording preferences"); GOOG-BRWN-00029433,

or former employee who Plaintiffs asked about it in depositions.[38] Mr. Hochman also recognized that Plaintiffs' proposed fingerprinting methodology would violate these policies.[39]

52. In addition to policies, Google maintains technical restrictions that further reduce the risk that the fingerprinting information Mr. Thompson identifies will be used to join signed-out private browsing data with users' Google accounts. For example, Google abrogates IP addresses *See* Rebuttal Report ¶¶ 118-19. In my opinion, these policy and technical restrictions further undermine Mr. Thompson's claim that it is easy to join data, as any joining of the data via the IP + UA fingerprinting method Mr. Thompson proposes would violate them. By contrast, for the reasons stated above, I conclude that merely adding records to the same log does not violate these policies and restrictions because merely adding records to the same log without associating any records does

at -435 ("You must not fingerprint users for the purpose of associating a user's activity over time or across contexts ('tracking') or re-identifying them when you do not have access to actual identifiers such as cookies or GAIA IDs."); GOOG-CABR-00086797, at -797 ("Device fingerprinting and immutable IDs policy: widely used across the company in design decisions (background input shows complexity) . . . Strengthens our position against market pressures to use fingerprinting for signed-out advertising and our position with regulators.").

[38] *See, e.g.*, McClelland Tr. 279:3-11 ("Q. Google also prohibits fingerprinting users for the purpose of associating their activity over time or across contexts, is that right? . . . A. Yes, that is my understanding, that fingerprinting was also not allowed to be used."); *id*. at 303:10-304:3 ("Q. Based on your work as product lead for Chrome browser privacy, would you agree that Google identifies users with fingerprinting techniques? A. I know that it was not allowed by policy and I never saw any evidence of it happening either. Q. Does Google use fingerprinting to identify users and personalize advertising, to your knowledge? A. I believe there may be some legitimate uses for fingerprinting around anti-fraud with sign-in, but understanding, again, is that it's not used for ad targeting and, again, I never saw any evidence to counter that."); Adkins Apr. 14, 2021 Tr. 188:5-10 ("[T]he Google services do not use any combination of identifiers to try to uniquely identify users, other than the Google logged in cookie and so, therefore, it's impossible because we don't conduct -- that's against our privacy policy."); id. at 314:2-8 ("[N]ot only is the practice discouraged or forbidden, but I am aware of there are active measures taken to prevent teams from being able to do fingerprinting within Google, so that it's not accidentally done, so fingerprinting is not accidentally done."); Bindra Feb. 8, 2022 Tr. 123:1-3 ("Q. So Google does not engage in fingerprinting; is that what you're saying? A. Correct."); Halavati Jan. 18, 2022 Tr. 151:2-5 ("Google has . . . strong regulations against fingerprinting and all -- all user identifiable data that is collected should be by direct user consent or signing into a Google account."); Monsees Apr. 9, 2021 Tr. 314:23-315:3 ("Q. So why is it Google's policy that you must not fingerprint users for the purpose of associating a user's activity over time or across contexts tracking? A. That would violate our policies and, I think, statements to our users and regulators."); Shelton Mar. 2, 2022 Tr. 141:10-16 ("Q. When you were working at Google, to your knowledge, was Google engaged in fingerprinting? . . . [A.] I'm not aware of Google doing what I have characterized here as fingerprinting.").

[39] Hochman Vol. II Tr. 447:24-449:11 ("Q. Turn back to the page ending with the Bates designation 006. We were talking about the 'Re-Identifying logs data' section [of Google's logs usage policy] . . . And in particular, the first bullet that I read out earlier in red that begins with 'you must not re-identify.' Do you see that? A. Yes, I do see that Google recognizes the danger of reidentification. Q. And my question to you, Mr. Hochman, is what is reidentify? A. Reidentify. So I'll give you a quick summary of it. If you have a record that looks like it's anonymous data, reidentifying means figuring out specifically who that data is associated with or what entity that data is associated with. Q. Would taking logged-out private browsing data at issue in this case and then using that to reidentify individuals who may have been associated with the browsing be reidentification in your opinion? A. Sure. If you take some logged-out browsing data and you figure out, for example, the GAIA ID of that person, or you -- you take the -- the IP address and user agent, which essentially is identifying, and you -- you track down the person, I think if you've identified someone based on an otherwise anonymous record, it's reidentification.").

not constitute "fingerprinting."

Opinion 3: Mr. Bhatia's Opinion That I Did Not Have A "Reasonable Basis To Say, 'The Coding Of The [[redacted]] Log Prohibits It From Ever Joining Authenticated With Unauthenticated Data,'" Bhatia Decl. ¶ 26, Is Without Basis.

53. Mr. Bhatia opines that "based on the source code that was made available, it does not appear that Google and Dr. Psounis had a reasonable basis to say, 'the coding of the log prohibits it from ever joining authenticated data with unauthenticated data—or, indeed, joining any records at all.'" Bhatia Decl. ¶ 26. He also asserts that "If [he is] being asked to reconcile Google's (and Dr. Psounis's) contradictory position [regarding purported joining of logs in the [redacted] log], they seem to be saying that the data is not 'joined' by a common key (e.g., GAIA), although the records are merged sequentially by time. If that is true, the debate is one of semantics." *Id.* ¶ 27. Mr. Bhatia also claims that Google's production of source code for his inspection is deficient because "the production did not contain definition source code for specific functions that are used to generate the [redacted] log," *id.* ¶ 17, "Google did not produce any source code that processes these other six logs [that are the input logs for the [redacted] log]." *id.* ¶ 18, and "the source code produced is incomplete because it omits the source code for several [redacted] directives, which the produced source code directly relies on," *id.* ¶ 19.

54. As explained below, these assertions are incorrect because (i) Mr. Bhatia's understanding of the term "joining" runs counter to how that term is commonly used in data analysis and computer science; (ii) Google produced all of the source code and technical documentation needed for me to form the opinions stated in my November 30, 2022 Declaration; and (iii) the records contained in the [redacted] log are not merged sequentially by time as the term "merged" is commonly understood in data analysis and computer science.

A. Mr. Bhatia's Understanding of "Joining" Also Runs Counter to Its Common Usage in Data Analysis and Computer Science, Along With More Than 70 Years of Technical Research

55. Mr. Bhatia asserts that Google has "seem[ingly] contradict[ed]" itself by explaining that (i) [redacted] contains records from two different sources;

but (ii) it does not join those records. Bhatia Decl. ¶ 10.

56. Mr. Bhatia does not provide a basis for his opinion that this language "seem[s] contradictory," but it appears to be attributable to an erroneous understanding of "joining" as that term is used in academia and in the field. *See supra* Opinion 1. When one applies the common and well-established definition of "joining" or "linking," it is no contradiction for any dataset to contain records from multiple sources without those records being *joined*. Based on my analysis of the source code underlying this log described in my November 30, 2022 Declaration, that is precisely what happens here (*i.e.*, [redacted] contains records from multiple input logs, but those records are not "joined" as that term is understood in this field because they are not combined together to "join" or "link" authenticated and unauthenticated data).

B. The Source Code That I Reviewed Provides a Sufficient Basis to Conclude That The Coding of the [redacted] Log Prohibits It From Ever Joining Authenticated Data With Authenticated Data.

57. Mr. Bhatia further opines that "based on the source code that was made available, it does not appear that Google and Dr. Psounis had a reasonable basis to say, 'the coding of the log prohibits it from ever joining authenticated data with unauthenticated data—or, indeed, joining any records at all,'" Bhatia Decl. ¶ 26.

58. In my opinion, Mr. Bhatia is incorrect because the materials I considered in reaching the conclusions in my November 30, 2022 Declaration are more than sufficient to show that the coding of [redacted] "prohibits it from ever joining authenticated data with unauthenticated data–or indeed, joining any records at all." Mr. Bhatia makes three claims to support his opinion, all of which are incorrect and without basis.

59. First, Mr. Bhatia claims that the source code Google produced was deficient because "the production did not contain definition source code for specific functions that are used to generate the [redacted] log." Bhatia Decl. ¶ 15. According to Mr. Bhatia, this means that "Google did not make available the basic source code that shows how the log is generated to begin with." *Id.* ¶ 17. That is not a deficiency that would impact any of the opinions expressed in my November 30, 2022 Declaration or Rebuttal Report because I reviewed the source code file ([redacted]) that contains the instructions for sorting input logs data in

the [redacted] log. As I explained in my November 30, 2022 Declaration, this file governs the steps for adding inputs to generate this log, which is the "specific function" responsible for "generating" the log. Thus, Mr. Bhatia's assertions are incorrect.

60. Second, Mr. Bhatia asserts that Google's source code production was deficient because "while Mr. Panferov discusses in his declaration how the [redacted] log is comprised of data from [redacted] other logs, Google did not produce any source code that processes these other [redacted] logs." Bhatia Decl. ¶ 18. But "[S]ource code that processes" the input logs is not necessary to determine whether or not joining of authenticated and unauthenticated data takes place. It is undisputed that each of the input logs either solely contain authenticated data or solely contain unauthenticated data, and thus any source code operating on one of the input logs cannot possibly join any authenticated data with unauthenticated data. November 30, 2022 Decl. ¶¶ 12–21. By contrast, the source code I reviewed confirms that the [redacted] log is generated by extracting the authenticated or unauthenticated identifier of each record, assigning these a log key value, and then *sorting* the records by the log key value one by one without *joining* any two records together. *Id.* ¶¶ 10–21. Mr. Bhatia does not appear to dispute this conclusion or provide his own analysis of the source code that Google made available for his inspection.

61. Third, Mr. Bhatia also opines that he "would have expected Dr. Psounis to identify the same deficiencies that I have identified and to request the same missing source code." Bhatia Decl. ¶ 20. In my opinion, this expectation is unreasonable because it is common for large code bases to contain a number of interdependencies in the form of [redacted] files,[40] and it is not necessary to analyze every [redacted] file to understand a particular function of the code base. Mr. Bhatia does not identify which "[redacted] directives" he would need to opine on the functionality of the source code on which my November 30, 2022 Declaration is based, but in my opinion, none of the [redacted] directives are necessary to confirm whether or not the log joins data.

62. What Mr. Bhatia failed to do in his analysis is also telling. To rebut my statements on this issue, I would have expected Mr. Bhatia to review the source code and examine its function

[40] John Lakos, *Large-Scale C++ Software Design*, 1st Ed. (1996).

to determine whether or not records within [redacted] are indeed joined, including the code snippets that were reproduced in paragraphs 14, 15, and 16 of my November 30, 2022 Declaration. Specifically, if Mr. Bhatia disagreed with my conclusions, I would have expected him to opine on what exactly the source code that I cited in my November 30, 2022 Declaration is doing, which, in my opinion, can be determined in an undisputed manner by a person of ordinary skill in the art (POSITA) without requiring additional "definition source code" of unspecified functions or the [redacted] files that Mr. Bhatia references. In fact, Mr. Bhatia offers no opinion on the functioning of the code described in my November 30, 2022 Declaration.

C. Mr. Bhatia's Opinion That Records in [redacted] *Are Joined or Merged Sequentially by Time is Without Basis.*

63. Mr. Bhatia asserts that:

> If I am being asked to reconcile Google's (and Dr. Psounis's) contradictory position, they seem to be saying that the data is not "joined" by a common key (e.g., GAIA), although the records are merged sequentially by time. If that is true, the debate is one of semantics.
>
> Bhatia Decl. ¶ 27.

64. As discussed above, the definition of "joining" is not a matter of "semantics" because the definition that I applied in my November 30, 2022 declaration aligns with an extensive body of technical literature dating back to at least 1946. Based on my experience and training, "joining" or "linking" of records is a term of art in data science with a well-settled meaning (*i.e.*, "joining" or "linking" refers to combining two records together with a common key to produce a single combined record). As explained above, it is my opinion that the [redacted] log does not meet this definition.

65. Mr. Bhatia's assertion that "the records [in [redacted]] are merged sequentially by time" is also incorrect because any "merging" of records would similarly require combining records via a common key. Simply adding multiple records to a single database and sorting them by time—without combining the data

contained in those records—does not constitute "merging" of records as that term is used in technical literature and in the field.[41]

<u>Opinion 4: Plaintiffs' assertion that a one percent random sample is invalid, Dkt. 833-1 at 11-12, is without basis.</u>

66. Plaintiffs state "Google . . . botched the limited investigation it sought to undertake," because "Google did not even search fields in 99% of [redacted] log traffic, instead limiting its review to searching a table containing a random 1% of logged traffic." Dkt. 833-1 at 11-12. In my opinion, Plaintiffs confuse the percentage of a sample with its validity. A sample is not invalid merely because it is based on 1% of random sampling. Random sampling is a tool commonly used in data analysis to work with a subset of a larger population for analysis.[42] Random sampling conserves resources while allowing the researcher to draw valid conclusions about the larger population.[43] In a large population of, for example, billions of log records, a mere 1% random sample would allow for analysis of tens of millions of records chosen via an unbiased random draw, which is a common practice in the field of data analysis.[44]

1. Consider a log with 1 billion records and a sample consisting of 1% of these records, or, equivalently, ten million of these records. If one wishes to use the sample to determine properties of the log, *i.e.* whether the log contains records with certain fields, then a sample of that size could be a representative sample of the original population.[45] Based on decades-old analytical results in probability theory and statistics, the sample could be used instead of the original, larger population

[41] *See, e.g.*, Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage," Journal of Amer. Statistical Assoc. 64:328, https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf at 51 (Dec. 1969) ("The necessity for comparing the records contained in a file L_A with those in a file L_B in an effort to determine which pairs of records relate to the same population unit is one which arises in many contexts, most of which can be categorized as either (a) the construction or maintenance of a master file for a population, or (b) *merging two files in order to extend the amount of information available for population units represented in both files*." (emphasis added)).

[42] See Mendenhall, W., and Sincich, T., "Statistics for Engineering and the Sciences", 6th Ed., CRC Press (2016); William G. Cochran, "Sampling Techniques", 3d Ed., John Wiley & Sons (1977); Hogg, R. V., Tanis, E. A., and Zimmerman, D., "Probability and Statistical Inference", 10th Ed., Prentice Hall (2019).

[43] *Id.*

[44] *See* Efron, B., Hastie, T., "Computer Age Statistical Inference: Algorithms, Evidence, and Data Science", Cambridge University Press, 2016; Hogg, R. V., Tanis, E. A., and Zimmerman, D., "Probability and Statistical Inference", 10th edition, Prentice Hall, 2019; Arnold, T., Kane, M., Lewis, B., "A Computational Approach to Statistical Learning", CRC Press, 2019.

[45] See Mendenhall, W., and Sincich, T., "Statistics for Engineering and the Sciences", 6th Ed., CRC Press (2016); William G. Cochran, "Sampling Techniques", 3d Ed., John Wiley & Sons (1977); Hogg, R. V., Tanis, E. A., and Zimmerman, D., "Probability and Statistical Inference", 10th Ed., Prentice Hall (2019).

to draw accurate conclusions about the original population.[46] To understand this, suppose one observes one by one the sampled records to identify if any of them contains specific fields. For the sample to incorrectly categorize the log as a log with no records with the specific fields, every single one of the ten million sampled records must not contain the aforementioned fields while there must exist at least one non-sampled record that does contain the specific fields. However, the probability of this occurring is for all practical purposes zero, as long as the log contains records with the specific fields, the percentile of such records over all log records is non-negligible. For example, even if records with the specific fields are very rare and thus hard to catch by sampling—for example, if only one in every hundred thousand records contains the specific fields—the probability of an erroneous determination whether this log contains records with the specific fields equals 0.99999^10000000 or 3.7e-44.[47] To get a sense of how small that number is, one has to divide 3.7 by 1 followed by 44 zeros. Hence, a 1% sample of the original population can be representative and a useful tool to accurately determine whether a log contains records with certain specific fields or not.

I declare under penalty of perjury that the foregoing is true and correct.

Executed on the 10th day of February, 2023 at Irvine, California.

By: DocuSigned by:
Konstantinos Psounis
EDAADC53351F49E...

Konstantinos Psounis, PhD

[46] *See* Sheldon Ross, "Introduction to Probability Models," Academic Press, (10th ed. 2014); Mendenhall, W., and Sincich, T., "Statistics for Engineering and the Sciences", 6th edition, CRC Press, 2016; William G. Cochran, "Sampling Techniques", 3rd Edition, John Wiley & Sons, 1977; Patrick Billingsley, "Probability and Measure," Wiley, 3rd ed. 1995.

[47] Suppose that sampling is used to identify whether a log contains records with a specific field and the chance that a record contains this field is 1/n. Then, if one samples N times n records, e.g. 10 times n, the probability of an erroneous determination equals (1-1/n)^(nN), which, as n grows, converges to (1/e)^N with e=2.718 being the so-called exponential constant. It is now evident why sampling is so powerful. Merely sampling 10 times more samples than the inverse of the chance of the specific field occurring yields a probability of erroneous determination of merely 45 over a million. Sampling 100 times more samples than the inverse of the chance of the specific field occurring and the probability of erroneous determination becomes infinitesimally small. Specifically, it equals 3.7e-44, that is, 3.7 divided by 1 followed by 44 zeros.